3К

УДК 681.3.01

И.П. Кузнецов, Е.Б. Козеренко, А.Г. Мацкевич
Институт проблем информатики РАН, г. Москва, Россия
igor-kuz@mtu-net.ru, kozerenko@mail.ru

Принципы организации объектно-ориентированных систем обработки неформализованной информации

Рассматривается класс логико-аналитических систем, использующих специальные лингвистические процессоры и базы знаний (БЗ) для обработки потоков неформализованных документов с целью решения пользовательских задач. На первом этапе формализации текста документа извлекаются информационные объекты и связи, которые образуют структуры знаний и запоминаются в БЗ. На уровне БЗ организуются различные виды анализа и объектных поисков: поиск похожих объектов и ситуаций, поиск по связям и другие. Рассматриваются основные компоненты подобных систем, называемых объектно-ориентированными, их особенности при использовании в различных приложениях: при обработке криминальной информации, при автоматической формализации резюме (заявок на работу), в системах обработки СМИ с выделением террористических групп и их деяний.

Введение

Лавинообразный рост потока документов, получаемых пользователями через различные информационные каналы, требует новых решений для повышения эффективности поиска и анализа необходимой пользователям информации. Большая часть таких документов имеет вид текстов на естественном языке (ЕЯ). Во многих случаях человек не в силах прочитать и осмыслить даже малую часть того, что ему предлагается. Существующие средства во многих случаях могут оказать лишь ограниченную помощь пользователям. Полнотекстовые базы данных не решают проблемы, так как при работе с текстами на ЕЯ дают много шумов (лишних документов) и потерь. Причина этого – особенности русского языка: наличие словоформ и свободный порядок слов. При использовании реляционных БД требуется трудоемкая работа специально обученных людей по формализации текстов на ЕЯ для заполнения соответствующих таблиц. При больших потоках документов это сделать крайне трудно. В любом случае будут потери той информации, которая не учтена в рамках схем БД. Описанная ситуация является типичной для многих областей, имеющих дело с потоками информации в виде текстов на ЕЯ.

Следует отметить, что большинство пользователей – это люди, которые интересуются конкретными вопросами. Например, следователю важны фигуранты, их места жительства, телефоны, криминальные события, даты и др. Специалиста по кадрам интересуют организации, где человек работал, кем он работал и когда это было. Другие люди вылавливают из СМИ информацию о странах, влиятельных лицах, катастрофах и др. Здесь важны и связи: места работы с занимаемой должностью, экстремальной ситуации с ее временем и т.д. Будем называть интересующую пользователя конкретную информацию – ***информационными объектами***. Каждый пользователь (или класс пользователей) интересуется своими информационными объектами и связями между ними. Вся остальная информация является лишней и человек старается ее просто не замечать.

Перспективное направление в области информатики – это обработка документов на ЕЯ, которая должна учитывать, прежде всего, интересы конечного пользователя. Отсюда следует необходимость построения нового класса информационных систем, использующих специальные лингвистические процессоры и технологию ***баз знаний*** (БЗ). Такие процессоры необходимы для глубинной обработки текстов с выявлением информационных объектов и связей. На основе последних формируются структуры знаний, которые образуют БЗ. На уровне БЗ становится возможным более полно учитывать потребности пользователя – за счет организации различных видов поиска: поиска конкретных объектов, похожих объектов, поиск по связям и др. Такие виды поиска относятся к семантическим или объектным и осуществляются не на уровне слов или словоформ, а на уровне структур знаний из БЗ. Будем называть системы подобного типа ***объектно-ориентированными***.

Это направление начало активно развиваться за рубежом [1], [2]. В данной работе будет идти речь о проблемах построения, основных компонентах, структуре и приложениях объектно-ориентированных систем, разрабатываемых в ИПИ РАН [3-5].

1 Структура объектно-ориентированной системы

На протяжении последних 15 лет в ИПИ РАН были разработаны различные классы объектно-ориентированных систем (ООС) в рамках проектов ДИЕС, «Аналитик», «Криминал», «Лингвопроцессор» и др. Основные задачи системы ООС: сбор всей поступающей информации (документов на ЕЯ), ее автоматическая формализация и хранение, а также решение задач семантического (объектного) поиска и анализа [3-6].

Система ООС ориентирована на автоматическую обработку документов в тех областях, где имеют место:

– большие потоки информации;

– неформализованный характер поступающей информации (это тексты на ЕЯ);

– высокая трудоемкость формализации документов специально обученными людьми;

– необходимость исключить последствия недобросовестной работы людей при формализации документов.

Основой системы ООС является лингвистический процессор, который обеспечивает автоматический ввод документов и их формализацию. В результате из документов выделяются объекты и их связи, на основе которых формируются структуры знаний, где содержатся только слова в нормальной форме.

1.1 Представление знаний

При разработке систем с БЗ важным фактором является выбор средств представления и обработки знаний. Наиболее адекватным средством представления и формализации знаний, выражаемых на ЕЯ, являются ***семантические сети*** следующего вида.

Семантическая сеть состоит из множества вершин, представляющих объекты. Из вершин составляются элементарные фрагменты, каждый из которых представляет k-местное отношение. В этот фрагмент вводится две дополнительных вершины: одна соответствует отношению, а другая (код фрагмента) – всей совокупности упомянутых объектов с учетом их отношения. Эти вершины, как и любые другие вершины, могут стоять на местах объектов в других фрагментах, что обеспечивает высокие изобразительные возможности и гибкость: представление отношений между отношениями, между совокупностями связанных объектов и т.д. [5], [7].

Множество вершин делится на два подмножества: первое соответствует распознанным или определенным компонентам (именам, понятиям), а второе – неопределенным объектам, т.е. вопросительным словам, различного рода умолчаниям. Последние играют роль переменных.

Из элементарных фрагментов составляются сети, называемые ***расширенными семантическими сетями*** (РСС). Как показали исследования, подобные сети оказываются удобными для представления семантической компоненты различных языковых конструкций, в том числе с отглагольными существительными и их формами, причастными оборотами, безглагольными конструкциями со связками типа «*это, есть, значит*» и др. Сети РСС служат в системах ООС для представления знаний. Для обработки структур знаний разработан ***инструментальный комплекс ДЕКЛ***, основой которого являются правила *ЕСЛИ…, ТО…,* осуществляющие преобразование сетей [5], [8].

Структуры знаний, представляющие формализованные документы, записываются в нотации расширенных семантических сетей – РСС, обладающих средствами представления безымянных объектов, событийных компонент и различного вида связей. В результате образуются так называемые ***содержательные портреты документов***.

1.2 Лингвистический процессор

Для построения содержательных портретов (т.е. структур знаний) используется лингвистический процессор, который включает в себя лексико-морфологический и синтактико-семантический анализ. За счет первого обеспечивается нормализация элементов текста (приведение словоформ к одному виду, что очень важно для поиска) и формирование признаков слов – лексических, морфологических, семантических [9]. За счет второго – автоматическое выделение из документа всей значимой информации: объектов и их связей [3], [5]. При этом «связанность» понимается в широком смысле. Несколько объектов, участвующих в одном действии, считаются связанными.

Особенности объектно-ориентированного ЛП состоят в следующем:

– поддержка модели языка с учетом семантических характеристик слов и словообразующих компонент;

– морфологический анализ слов с учетом приставок, словообразующих суффиксов и отглагольных форм;

– синтаксический и семантический анализ текстов, выделение объектов, их признаков и связей с автоматическим формированием структур знаний – в виде РСС;

– наличие предметных словарей и родовидовых деревьев (онтологий), используемых для семантического анализа текстов;

– анализ анафорических ссылок (местоимений) с идентификацией соответствующих объектов;

– выделение признаков, связей, относящихся к описываемому значимому объекту, сбор сведений об объекте;

– восстановление информации об объектах и связях, данной в неявном виде.

При разработке ЛП удалось преодолеть многие трудности, связанные с множественностью форм выражения и неоднозначностями на различных уровнях анализа компонентов текста на ЕЯ. Например, на уровне словоформ необходимо учитывать словообразующие суффиксы, не изменяющие смысла слова и используемые для поддержания соответствующих языковых форм, например, *бородатый, бородатые, с бородой* и т.д. Далее, приходится учитывать термины различного уровня общности, например, *пистолет, огнестрельное оружие*, а также случаи омонимии существительных и полисемии глаголов. Такое разнообразие учитывается путем использования в лингвистических знаниях синонимичных рядов, терминов, родовидовых или SUB-деревьев (в рамках

онтологий). Здесь большую роль играет контекст. Например, *организация* – это может быть действие, а может быть и юридическое лицо. Особое место занимает расшифровка сокращений – путем анализа контекста. На ЕЯ одно и то же действие можно выразить по-разному: с помощью глагольных форм, отглагольных существительных, причастных оборотов и др. Объектно-ориентированный ЛП обеспечивает их представление в БЗ с помощью одних и тех же структур знаний.

1.3 Принципы обработки

Система ООС содержит собственную базу данных, которая служит для хранения поступающих документов и сформированных структур знаний. Последние образуют долговременную **базу знаний** (БЗ). При этом из формализованных документов (структур знаний) автоматически выделяются ключевые слова. На их основе строятся предметные каталоги и индексные файлы, обеспечивающие быстрый выбор из долговременной БЗ необходимых структур знаний с созданием в оперативной памяти оперативной БЗ, которая служит основой для поиска и решения прикладных задач. Это осуществляется следующим образом.

Пусть на вход системы поступил запрос на ЕЯ с требованием найти какой-либо объект. Запрос формализуется с выделением объектов и их связей. Образуется структура знаний, где все слова приведены в нормальную форму. Из них выделяются значимые слова, которые *характеризуют* объект. По индексным спискам находятся документы, содержащие такие же слова или их подмножество. По степени значимости совпавших слов подсчитывается *вес* каждого найденного документа. Содержательные портреты документов с наибольшими весами считываются в оперативную память и образуют оперативную БЗ. Далее начинается поиск требуемого объекта – путем сопоставления структур, представляющих запрос, и оперативных знаний. В рамках систем ООС реализованы различные объектные поиски, среди которых следует выделить: точный поиск объекта, поиск похожих, поиск по связям (приметам), поиск связанных объектов и др. Опыт показывает, что при такой организации потери информации минимальны. Аналогичным образом идет поиск нескольких объектов, ответ на запросы в формах ЕЯ, реализация логико-аналитических функций, где идет постоянное обращение к поисковым процедурам (п. 2).

Рассмотрим более подробно особенности систем ООС для различных областей приложения.

2 Система «Криминал»

Потоки документов в криминальной милиции – это сводки происшествий, справки по уголовным делам, обвинительные заключения и др. В этих документах содержится много конкретной информации, касающейся фигурантов, их деяний, орудий преступления и др. Основные задачи – различные виды поиска и логико-аналитическая обработка. Отметим, что объемы ежемесячной новой информации подобного типа исчисляются десятками и сотнями мегабайт. Никто не может все это прочитать и держать в голове. Как уже говорилось, использование БД создает определенные трудности при решении многих задач следователей-аналитиков.

2.1 Особенности системы «Криминал»

В связи с этим в конце 90-х годов в рамках проектов ООС была разработана система «Криминал» [3], [5]. Ее особенность – автоматический анализ текстов с выделением необходимого набора информационных объектов. Система «Криминал» отла-

живалась на 500 тыс. происшествий из сводок ГУВД г. Москва и по основным объектам удалось добиться хороших результатов: коэффициент шумов в компонентах (лишних слов в объектах) – не более 1 – 2% и потерь (отсутствие нужных слов) – не более 1%.

Основные выделяемые объекты (потери должны быть минимальными):

– лица (по ФИО) с их особенностями (преступник, потерпевший);
– словесное описание лиц, их приметы;
– адреса, почтовые атрибуты;
– даты;
– оружие с атрибутами;
– номера телефонов, факсов, е-майлов с их стандартизацией;
– средства транспорта с выделением марки машины, государственного номера, цвета и других атрибутов;
– паспортные данные и другие документы с их атрибутами;
– взрывчатые вещества и наркотические вещества;
– отделения милиции;
– сотрудники милиции.

Второстепенные объекты (потери допустимы):

– организации;
– должности;
– количественные характеристики (сколько лиц или других объектов принимали участие в том или ином событии);
– номера счетов, суммы денег с указанием типа валюты;

Связи:

– события (криминальные, террористические, поломки изделий и др.) с указанием участия в них информационных объектов;
– время и место событий;
– связи между различными типами информационных объектов (кем работает лицо в той или иной организации, по какому адресу проживает, в каких событиях принимал участие совместно с другими объектами и т.д.).

Особенности текстов в области «Криминалистика» это, во-первых, наличие (особенно в сводках происшествий) большого количества сокращений, которые нужно расшифровывать путем анализа контекста. Например, *Г.* может означать *ГОД, ГОРОД, ГОС.* и др. Во-вторых, много подразумеваемой информации. В наибольшей степени это относится к связям. Например, после фигуранта пишется его адрес, год рождения и другие данные. Их нужно связывать с фигурантом. Еще одна не простая задача – идентификация объектов (фигурантов) по всему тексту, использование для этих целей указательных местоимений, кратких имен, анафорических ссылок. Это особенно необходимо для обвинительных заключений, где одно и то же лицо упоминается многократно (различными способами именования) по всему документу.

С учетом трудностей и в соответствии с задачами был разработан лингвистический процессор системы «Криминал», осуществляющий нормализацию слов, их группировку с формированием объектов, идентификацию объектов и установление связей. В результате по каждому документу ЕЯ автоматически формируется структура знаний – содержательный портрет документа. Такие структуры запоминаются в БЗ, на основе которой реализованы различные виды семантического поиска: поиск по признакам и связям, поиск связанных объектов на различных уровнях, поиск похожих фигурантов и происшествий, поиск по приметам (с использованием онтологий).

Поддерживается **экспертная компонента.** Например, для классификации происшествий по каталогам криминальной милиции: «Вид преступления», «Способ совершения преступления» и др. Результат вводится в содержательный портрет.

2.2 Пример содержательного портрета

Пусть имеется следующий текстовый документ:

24. *Обман потребителей и задержание* *Западное ОУВД ОМ мо «Филевский парк»*

25.05.98г. в 16.40 уч. инспектором Маркиным на рынке по адресу:
ул. Барклая, 10 была задержана Сивушева Ольга Николаевна, 1965г.р., прож. Сеславинская 30-25, продавец ТОО «Ника», которая совершила обман троих покупателей на сумму 14 руб.
Подписка о невыезде. Дозн. Федосейкин.

Содержательный портрет этого документа имеет следующий вид:
ДОК_(24,1-96.TXT, "Сводка;")
ОВД_(ЗАПАДНЫЙ,ОУВД/1+) DESC_(1-,"Западное ОУВД ",39)
ОВД_(ОМ,МО,ФИЛЕВСКИЙ,ПАРК/2+) DESC_(2-,"ОМ мо ` Филевский парк ` ",93)
ЗАДЕРЖАТЬ(2-/3+) DESC_(3-,"задержание ОМ мо ` Филевский парк ` ",59)
ДАТА_(1998,05,25,16.40/4+) DESC_(4-,"25.05.98. в 16.40 ",133)
МИЛ_(ИНСП.,МАРКИНЫМ/5+) DESC_(5-,"инспектор Маркин ",156)
FIO(СИВУШЕВА,ОЛЬГА,НИКОЛАЕВНА,1965/6+)
DESC_(6-,"Сивушева Ольга Николаевна , 1965 год р. ",235) DESC_(6-,"которая ",326)
АДР_(СЕСЛАВИНСКАЯ,30,25/7+) DESC_(7-,"прож. Сеславинская 30 - 25 ",279)
ПРОЖ.(6-,7-)
ЗАДЕРЖАТЬ(6-/8+) DESC_(8-,"задержана Сивушева Ольга Николаевна , 1965 год р. ",186)
АДР_(УЛ.,БАРКЛАЯ,10/9+) DESC_(9-,"адрес : ул. Барклая , 10 ",189)
Где(8-,9-) Где(8-,РЫНОК)
ОРГ_(ТОО,НИКА/10+) DESC_(10-,"ТОО Ника ",314)
РАБ_(6-,ПРОДАВЕЦ,10-/11+) DESC_(11-,"продавец ТОО Ника ",305)
КОЛИЧ_(3,ПОКУПАТЕЛЬ/12+) DESC_(12-,"трое покупателей ",358)
КОЛИЧ_(СУММА,14,РУБ./13+) DESC_(13-,"сумма 14 руб.",379)
ОБМАН(12-,НА,13-/14+) DESC_(14-,"обман троих покупателей на сумму 14 руб.",344)
СОВЕРШИТЬ(14-/15+) DESC_(15-,"совершила обман троих покупателей на сумму 14 руб.",334)
МИЛ_(ДОЗНАВАТЕЛЬ,ФЕДОСЕЙКИН/16+) DESC_(16-,"Дозн. Федосейкин ",431)

ПРЕДЛ_(0,п.23,ОБМАН,ПОТРЕБИТЕЛЬ,И,1-,3-/17+) 17-(1,2,133)
ПРЕДЛ_(0,4-,УЧ.,5-,8-,7-,11-,6-,15-/18+) 18-(3,134,410)
ПРЕДЛ_(0,ПОДПИСКА,О,НЕВЫЕЗД/19+) 19-(7,411,431)
ПРЕДЛ_(0,16-/20+) 20-(7,432,447)
ANAL_("Преступные действия",МОШЕННИЧЕСТВО)

Фрагмент ДОК_(24,1-96.TXT, "Сводка;") указывает на порядковый номер документа (24-й) и имя файла 1-96.TXT, содержащего сводку с данным документом.

Фрагменты ОВД_(ЗАПАДНЫЙ,ОУВД/1+) DESC_(1-, "Западное ОУВД", 39) представляют «отделение внутренних дел» с его описанием DESC_, взятое из текста с указанием месторасположения в байтах – 39. Такие описания даются для всех выделенных объектов (действие или событие тоже считается объектом). Коды 1+ (это код фрагмента) и 1 – указывают, что описание относится к данному ОВД_. Фрагмент FIO(СИВУШЕВА,ОЛЬГА,НИКОЛАЕВНА,1965/6+) представляет фигуранта с ФИО

и годом рождения. Фрагмент с именем МИЛ_ представляет «сотрудников милиции», ДАТА_ – «дату», АДР_ – «адрес» и т.д. Фрагмент ПРОЖ.(6-,7-) представляет, что *фигурант проживает по адресу* АДР_(…./7+).

Фрагменты: ЗАДЕРЖАТЬ(6-/8+)АДР_(УЛ.,БАРКЛАЯ,10/9+) Где(8-,9-) Где(8-, РЫНОК) представляют действие, что *фигурант был задержан «на ул. Барклая, 10» и «на рынке».*

Фрагменты ПРЕДЛ_ представляют предложения с аргументами: кодами фрагментов, которые представляют объекты и действия, и словами, которые никуда не вошли. За счет фрагментов ПРЕДЛ_ и DESC_ текст может быть восстановлен по содержательному портрету документа. Наконец, последний фрагмент – аналитический, который порождается экспертной системой, относящей происшествие к определенному классу – МОШЕННИЧЕСТВО.

Подобные содержательные портреты являются удобным формализмом для многих задач:

– для организации различных видов поиска, так как все слова представлены в нормальной форме и сгруппированы по объектам и действиям;

– ответ на запросы в свободной форме (на ЕЯ);

– поиск связей между объектами;

– выявление и ранжирование объектов по качественным критериям, заданным пользователем (криминальная активность и др.);

– для построения различных классов экспертных систем (на языке ДЕКЛ, у которого основные типы данных – такого же сора фрагменты);

– для построения графических схем, протоколов, аннотаций, кратких описаний, отражающих особенности интересующих пользователя объектов (за счет фрагментов DESC_);

− для заполнения таблиц и различных форм.

3 Задачи кадровых агентств

Одна из важных проблем кадровых агентств связана автоматической обработкой автобиографических данных, заявок на работу (резюме), написанных в произвольной форме – в виде текстов ЕЯ. Такие тексты содержат сведения о человеке: ФИО, год рождения, адрес, время и место учебы с указанием наименования учебного заведения и др. Требуется их автоматическая формализация с выделением информационных объектов и их отображением на поля заданной анкеты или сайта.

Тогда становится возможным использование типовых средств баз данных для решения пользовательских задач. Во многих агентствах такая формализация делается вручную: специально подготовленными людьми, или же самим человеком, которому предлагается ввести его сведения в указанные поля по требуемой форме. Это достаточно трудоемкая работа.

В качестве основы для автоматизации этих работ был взят лингвистический процессор системы «Криминал». Однако он был доработан в соответствии с особенностями предметной области [6]. Во-первых, это необходимость выделения другого набора объектов и связей. Во-вторых, их деление на группы. Например, деление объектов (организаций, дат и др.) на те, которые относятся к учебе или к профессиональной деятельности, или к курсам. В-третьих, необходимость использования экспертных систем для пополнения данных, которые заданы в неявном виде. Будем называть такие данные экспертными объектами.

Основные объекты:

ЛИЦО

– лицо, составляющее ЗАЯВКУ (как правило, в самом начале заявки); дата рождения или возраст; E-mail; почтовый адрес; домашний телефон; мобильный телефон; рабочий телефон; личная интернет-страница; желаемая должность;

УЧЕБА

– название учебного заведения; факультет (специальность); диплом (степень); начало учебы (дата); окончание учебы (дата);

ПРОФЕССИОНАЛЬНЫЙ ОПЫТ

– начало работы (дата); окончание работы (дата); название организации;

– занимаемая должность; обязанность, функции, достижения;

КУРСЫ (обучение)

– проводящая организация; название курсов; диплом (сертификат); начало курсов; окончание курсов.

Экспертные объекты:

– пол; образование (среднее, высшее и др.); профессиональная область (по заданной классификации); специализация (по заданной классификации); опыт работы (суммируется количество лет); регион (вычисляется по адресу); знание языка (по степени владения).

3.2 Особенности анализа

Выделение многих из этих объектов потребовало лишь доработки лингвистических знаний (ЛЗ). Однако особенности текстов и решаемые задачи потребовали усиления возможностей самого ЛП. Это было вызвано следующими факторами. Во-первых, разнообразием форм ЕЯ, с помощью которых выражаются даты и временные интервалы. Например, даты могут быть в сокращенной форме (*авг.05*), в виде дробных чисел (*09.99 г.)*, разного рода специальных знаков или кавычек (*09/99* или *09'1999*) и т.д. Интервалы: *15.05 – 01.12.99* или *май-июнь 06* и др. Трудности вызывали их путаница с дробными числами, отсутствие ключевых слов типа *г. (год)* и др. Более того, одним из требований было приведение дат к стандартному виду – расшифровка сокращений.

Во-вторых, определенные трудности вызывали задачи деления объектов на типы и правила их компоновки: необходимость выработать формальные критерии выявления, разделения и соотнесения дат, которые бы давали допустимое количество шумов и потерь. В связи с этим в ЛП были введены специальные средства, которые, опираясь на даты (или организации), осуществляли поиск связанных с ними объектов.

В-третьих, многие пользователи создавали свои резюме на основе документов, взятых из различных таблиц, форм. Как следствие, отсутствие знаков препинания (точек), наличие спецзнаков, остающихся после перекодировки текстов. Все резюме (если не было пробельных строк) воспринималось как одно предложение.

В связи с этим в блок лексико-морфологического анализа были введены специальные средства настройки – правила для выделения предложений [9]. Например, если слово-глагол написано с большой буквы и стоит вначале строки, то это начало предложения. Таких правил множество, в том числе такие, которые учитывают роль спецзнаков, разделительных символов и др.

В-четвертых, для получения экспертных данных (объектов) в ЛП были встроены экспертные системы (ЭС), которые, например, на основе анализа содержательных портретов соотносят документ к определенной категории (пункту классификатора), или же на основе имеющегося описания определяют степень владения иностранным языком и т.д. Если такая информация указана в исходном тексте в явном виде, то экспертной оценки не требуется.

В системе реализовано два типа оболочек для ЭС. Первая основана на весовых коэффициентах слов, соответствующих определенной категории. Вторая – на наличии слов в информационных объектах.

В ЭС первого типа с каждой категорией связываются слова с указанием их весов. Такие веса являются результатом статистического анализа эталонных документов (проанализированных человеком), т.е. предполагается этап обучения.

В ЭС второго типа с каждой категорией связываются характеризующие слова или пары слов (словосочетания), которые берутся из фрагментов, соответствующих информационным объектам указанного типа. Одно и то же слово или словосочетание может соотноситься лишь с одной категорией.

И, наконец, последнее – это необходимость в обратном ЛП. Обратный ЛП служит для преобразования объектов в компоненты ЕЯ и для их отображения на поля анкеты или сайта. Этот процессор имеет свои лингвистические знания, с помощью которых задается последовательность выдачи рубрик (полей) и какими объектами они должны заполняться. Для выделения таких объектов служат их имена (ОРГ_, РАБ_,...), а также связи, заданные в содержательном портрете. Для каждого выделенного объекта строится его описание – из входящих в него нормализованных слов. Далее, по объекту находится его предложение. За счет средств позиционирования находится место предложения в тексте, т.е. интервал от байта до байта. По описанию объекта в этом интервале ищется кусок предложения, соответствующий объекту. Этот кусок и выдается в качестве результата.

3.3 Пример разбора

Пример разбора текста резюме приведен в табл. 1.

Таблица 1 – Разбор текста резюме

РЕЗЮМЕ на ЕЯ:	В результате обработки данного РЕЗЮМЕ автоматически порождается следующая ФОРМА:
Иванова Таисия Петровна *+7(910)412-99-57(моб.),* *692-57-62 (дом.)* *Ivanova_tp@mail.ru* *35 лет.* *Образование Московский Государственный Строительный Университет* *Специальность: Инженер-строитель-технолог* *Опыт работы:* *2003–2005 ООО «Бизнесцентр Жемчужный»*	Язык резюме:-> Русский Фамилия:-> *Иванова* Имя:-> *Таисия* Отчество:-> *Петровна* Пол:-> Женский Дата рождения:-> Возраст:-> *35* Образование:-> ## 1 - (Высшее) Желаемая должность:-> Профессиональная область:->##19 - Строительство. Вес 27.11 Специализация:-> Опыт работы:-> ## 1 - (1-3 года) Ожидаемый уровень месячного дохода:-> *От 800 евро* Регион:-> Возможность переезда:-> E-mail:-> *ivanova_tp@mail.ru* Почтовый адрес:-> Домашний телефон:-> *692-57-62* Мобильный телефон:-> *+7(910)412-99-57* Рабочий телефон:-> Телефон:-> Личная интернет-страница:->

Продолж. табл. 1

Начальник отдела аренды нежилых помещений *Руководство отделом (5 человек), ведение переговоров* *Заработная плата от 800 у.е.* . . .	УЧЕБА:-> Название:-> *Московский Государственный Строительный Университет* Факультет (специальность):-> *Инженер строитель технолог* Диплом (степень):-> Начало учебы:-> Окончание учебы:-> ПРОФЕССИОНАЛЬНЫЙ ОПЫТ:-> Начало работы:-> *2003* Окончание работы:-> *2005* Название организации:-> *ООО Бизнесцентр Жемчужный* Занимаемая должность:-> *Руководство отделом. Начальник отдела аренды нежилых помещений* Обязанность, функции, достижения:-> *Руководство отделом (5 человек),* *ведение переговоров*

Другое приложение системы ООС это анализ текстов, выявление объектов и заполнение ими полей БД.

4 Документы СМИ о террористической деятельности

Основная задача – выделение из потока сообщений СМИ тех документов, которые относятся к террористической деятельности, с последующим анализом этих документов [4], [10]. В качестве основы служила система «Криминал». Лингвистический процессор (ЛП) этой системы был доработан в соответствии с особенностями предметной области и задач. В ЛП были дополнительно введены следующие информационные объекты:

– террористические группы и организации (Terrorizm);

– участник террористические группы с указанием его роли (лидер, главарь и др.);

– вооруженные силы, используемые для борьбы с терроризмом (Military_Force);

– интервалы времени (п. 3).

Были разработаны лингвистические знания (ЛЗ) для выделения этих объектов. В соответствии со спецификой текстов ЛЗ были дополнены новыми правилами выделения объектов, например, выделение места события по формам «*в 25 км. от Кабула»* или «*лагерь близ города Умма*» и т.д. Особые трудности вызывало выделение арабских составных имен с их элементами *абд* (раб), *Абу* (отец), *Ибн* или *Бен* (сын) и др. Они не укладываются в формат европейских стандартов. Например, *Абд ар-Расул бен-Ахмад*. Соответственно, усложняется ФИО. Для известных террористов, как правило, используются сокращенные имена, например, *Бен Ладен* (вместо *Усама Бен Ладен*), *Басаев (Шамиль Басаев)* и др. В ЛП были введены специальные средства их идентификации.

Как и в предыдущих случаях, при выделении объектов учитываются возможные варианты называния объекта в тексте, в том числе краткой форме. Типовые объекты (ФИО, даты, адреса, виды оружия и др.) приводятся к одному (стандартному) виду. Осуществляется идентификация объектов с учетом кратких наименований (например, отдельных фамилий или имен с ФИО), анафорических ссылок (указательных и личных местоимений, например, «*Этот человек», «Он ...*»), определений (например, «*Мэр Москвы Лужков»* идентифицируется с последующими словами «*мэр», «Лужков*"). Для выделения событий и связей проводится анализ глагольных форм, а также причастных и деепричастных оборотов.

В результате строились содержательные портреты, которые запоминались в долговременной БЗ. На их основе решались те же задачи, что и в системе «Криминал»: организация различных видов поиска, ответ на запросы, выраженные на ЕЯ, формирование дополнительных признаков у объектов (террористов), выявление их связей и др. Для решения были разработаны программы на языке ДЕКЛ, осуществляющие соответствующие преобразования структур знаний.

Заключение

Объектно-ориентированные системы обработки неформализованной информации, представленной в виде текстов на естественном языке, – это перспективное направление с широким кругом приложений. Интерес к такого сорта системам неизменно растет. Основное их назначение – это анализ потока сообщений, их автоматическая формализация с накоплением в базе знаний и последующим использованием для постоянного информирования пользователя в его предметной области. Такие системы находят свое применение для дифференцированного сбора информации (в том числе – из сети Интернет), выделения из нее интересующих пользователя объектов с их анализом и выдачей пользователю результатов в наиболее удобном в виде: протоколов, графических схем, форм с заполняемыми полями и др.

Литература

1. FASTUS:a Cascaded Finite-State Trasducerfor Extracting Information from Natural-Language Text // AIC, SRI International. – Menlo Park. California, 1996.
2. Кузнецов И.П. Семантико-ориентированные системы на основе баз знаний : [монография] / И.П. Кузнецов, А.Г. Мацкевич. – М. : МТУСИ, 2007. – 173 с.
3. Кузнецов И.П. Методы обработки сводок с выделением особенностей фигурантов и происшествий / И.П. Кузнецов // Труды Международного семинара «Диалог-1999» по компьютерной лингвистике и ее приложениям. – Таруса, 1999. – Том 2.
4. Kuznetsov I. The system for extracting semantic information from natural language texts / Kuznetsov I., Kozerenko E. // Proceeding of International Conference on Machine Learning. MLMTA-03, (Las Vegas US, 23-26 June 2003). – P. 75-80.
5. Kuznetsov I.P. Tools for Tuning the Semantic Processor to Application Areas / I.P. Kuznetsov, D.A. Efimov, E.B. Kozerenko // Proceedings of ICAI'09, WORLDCOMP'09, (July 13–16, 2009, Las Vegas, Nevada, USA) Vol. I. – Las Vegas : CRSEA Press, 2009. – P. 467-472.
6. Кузнецов И.П. Семантико-ориентированный лингвистический процессор для автоматической формализации автобиографических данных / И.П. Кузнецов, А.Г. Мацкевич // Труды Международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2006». – Бекасово, 2006. – С. 317-322.
7. Кузнецов И.П. Семантические представления / Кузнецов И.П. – М. : Наука, 1986. – 290 с.
8. Кузнецов И.П. Система обработки декларативных структур знаний ДЕКЛАР-2 / И.П. Кузнецов, М.М. Шарнин. – М. : ИПИАН, 1988.
9. Сомин Н.В. Система морфологического анализа: эксплуатации и модификации. / Н.В. Сомин, Н.С. Соловьева, М.М. Шарнин // Системы и средства информатики. – Вып. 15. – 2005. – С. 20-30.
10. Voss S. Advanced Knowledge Integration in Assessing Terrorist Threats / S. Voss, C.A Joslyn // LANL Technical Report LAUR 02-7867. – 2002.

Igor P. Kuznetsov, Elena B. Kozerenko and Andrew G. Matskevich

The Organization Principles of the Object-Oriented Systems for the Unstructured Text Information Processing

A class of the logical-analytical systems using special linguistic processors and knowledge bases is considered. Such systems are called object-oriented. These systems are employed for processing of the unstructured documents flow for the user problems decision. At the first stage the document text is formalized: information objects and links are extracted and transferred into the knowledge structures which are stored in the knowledge base (KB). At the level of KB various kinds of analysis and object search are organized: the search for similar objects and situations, the search on the basis of links and other types of search. The basic components of these systems, their main features and the particular use in different applications are considered.The system operation in the subject areas of criminal information processing, automatic formalization of summary texts (applications for work), mass media analysis for extracting information about terrorist formations and their activities are presented.

Статья поступила в редакцию 21.06.2008.